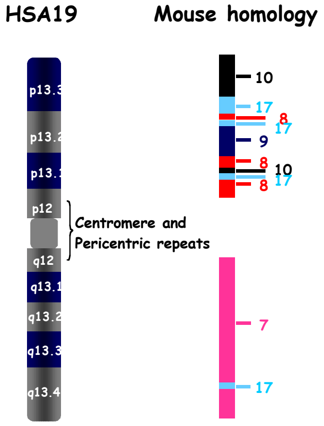
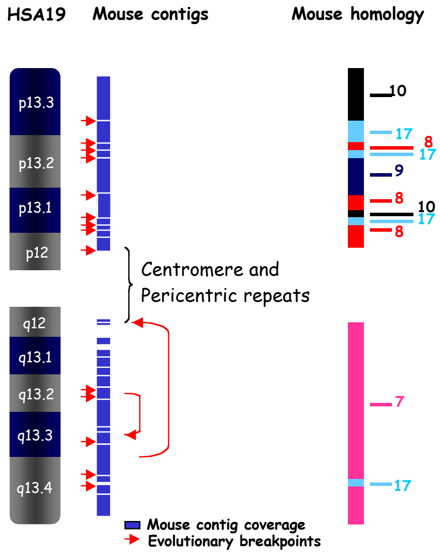
Chromosome 19:

- ~65-70 Mb total length
  - estimate of up to 1100 genes
  - ~17 Mb centromere + pericentromeric repeats (few or no genes)
  - ~2 Mb gene "deserts"
  - 46 Mb gene containing regions targeted for comparative sequencing

- 57 Mb contiguous clone map with 7 gaps
  - 35 Mb finished sequence
  - 22 Mb mostly o&oed draft

- 15 homology segments related to Mmu7, 8, 9, 10 and 17





Status

~35 mouse BAC contigs spanning the lengths of all 15 homology segments

- several homology segments spanned by a single contig
- > 95% coverage of mouse chromosome 19-related regions
- breakpoints of all evolutionary rearrangements cloned

>42 Mb non-overlapping mouse draft sequence completed

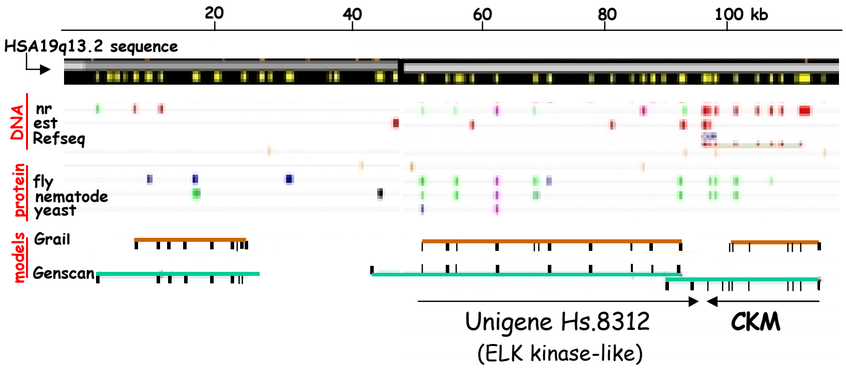
- All clones sequenced at ≥ 6X depth in paired plasmid ends
- Sequence of >60% clones is fully ordered and oriented; most remaining are partially ordered

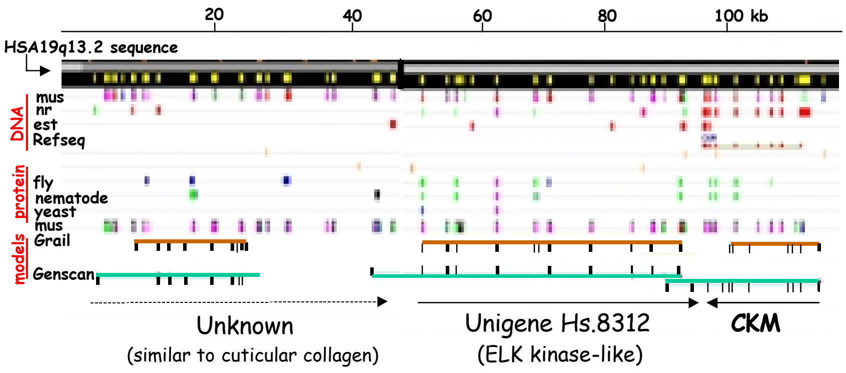
Initial analyses focused on three major questions:

- Human sequence annotation:
  - **value of comparative alignments for gene finding and functional-element definition**
- Chromosome evolution:
  - **what clues are provided by analysis of sequence at breaks in syntenic homology?**
- Gene evolution:
  - **How do primate and rodent gene sets compare?**
    - **What impact might species-specific differences have on biology?**

Value of comparative sequence alignment as a sequence-annotation strategy

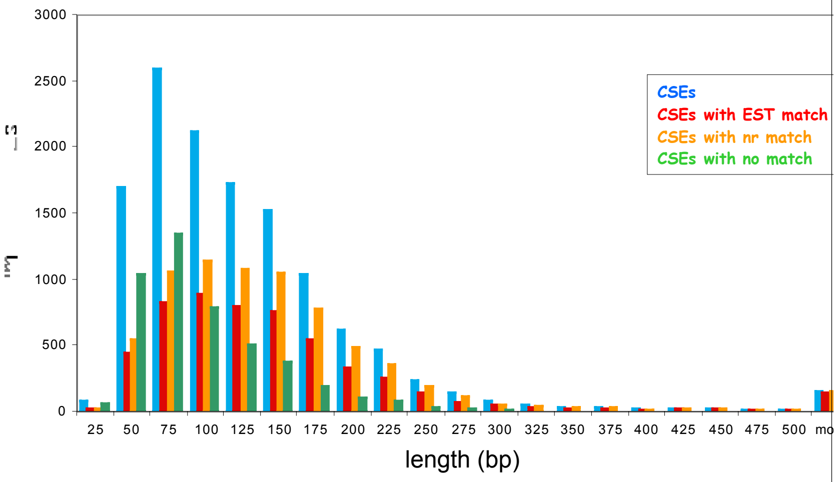
- What did we gain by sequencing mouse?
  - Identification of many new candidate exons -- 5′ ends, alternative exons, etc. in known genes
  - Confirmation and expansion of predicted genes
  - Prediction of ~30 new candidate genes that would have been missed entirely by other gene-finding methods
  - 128 genes identified by EST + mouse conservation only
  - >4000 non-coding conserved sequences that are candidates for regulatory DNA sequence elements





HSA19 conserved sequences

- 5.4% of HSA19 sequence is conserved at significant levels in syntenically homologous mouse DNA
  - aligning non-homologous mouse sequence yields more conserved elements, but most are *not* functionally significant
- 80% of the exons of known HSA19 genes are conserved in homologous regions of mouse
- 12611 conserved sequence elements:
  - 42% coincide with high-probability EST matches
  - 57% have significant match to non-redundant nucleotide or protein database entries
  - 36% conserved sequences (4546 CSEs) do not coincide with any other sequence feature (EST, protein, nt, exon model)



Gene number predictions and general observations

- Combining mouse sequence matches with other evidence we predict ~1200 HSA19 genes
  - ~860 (~70%) "unique" (or small-family) genes
  - ~340 (~30%) members of large clustered families
    - ZINC FINGER GENES, OLFACTORY RECEPTORS, VOMERONASAL RECEPTORS, CYTOCHROME P450 GENES, NATURAL KILLER RECEPTORS, SERINE PROTEASES, PREGNANCY SPECIFIC GLYCOPROTEINS, SIALIC ACID GYCOPROTEINS…...
- All but 30 predicted genes based on high probability EST matches
- Computer-based gene finding programs found one or more exons in ~55% of HSA19 genes (60% of known)

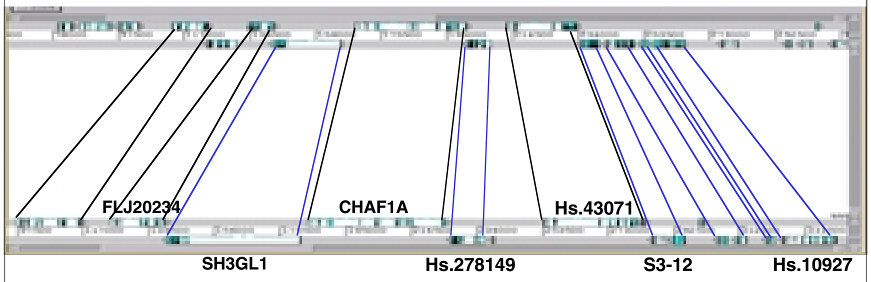
III. Conservation of human and mouse gene sets

Unique HSA19 genes are overwhelmingly conserved in mouse...

- Of 781 established (Refseq, Locus link, unigene) HSA19 genes, clear relatives were found for 744 (95%) in related mouse BAC sequence
  - 31 genes fall into gaps in the mouse BAC map (4.1%)
  - 3 genes are missing from well-covered mouse regions
    - PPPAR1A: member of gene distributed family
    - 2 unigene matches encoding hypothetical proteins

- In general, orthologous genes are arranged in identical order in homologous regions of HSA19 and mouse DNA…

Mouse Chromosome 17

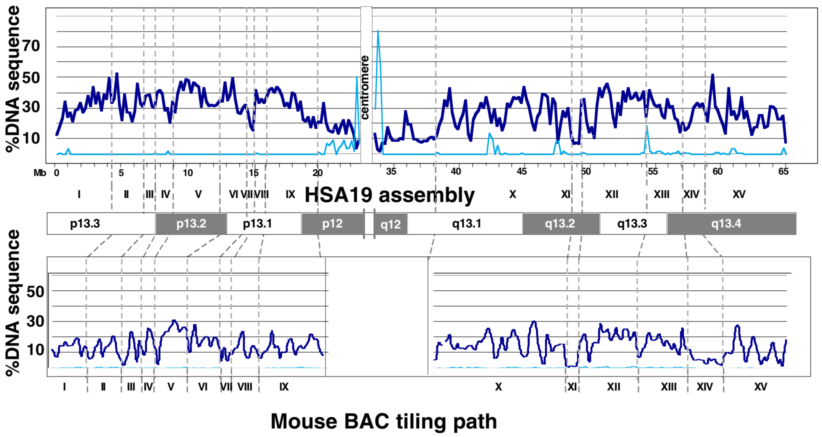


Human 19p13.3

….But in regions totaling ~50% of HSA19 DNA, human genes are larger and more widely spaced due to a significant increase in the numbers of inserted SINE repeats

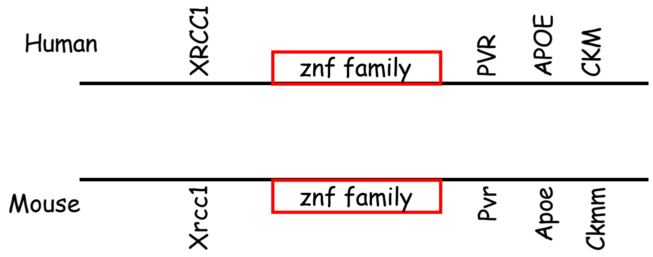
Sines as percentage of total sequence, plotted along HSA19 and mouse

HSA19 average: ~27%   mouse: 12.7%



...Tandem gene families

- MOST CLUSTERED HUMAN FAMILIES ARE REPRESENTED BY A RELATED FAMILY IN THE SYNTENICALLY CONSERVED POSITION


Human

XRCC1

znf family

PVR

APOE

CKM

Mouse

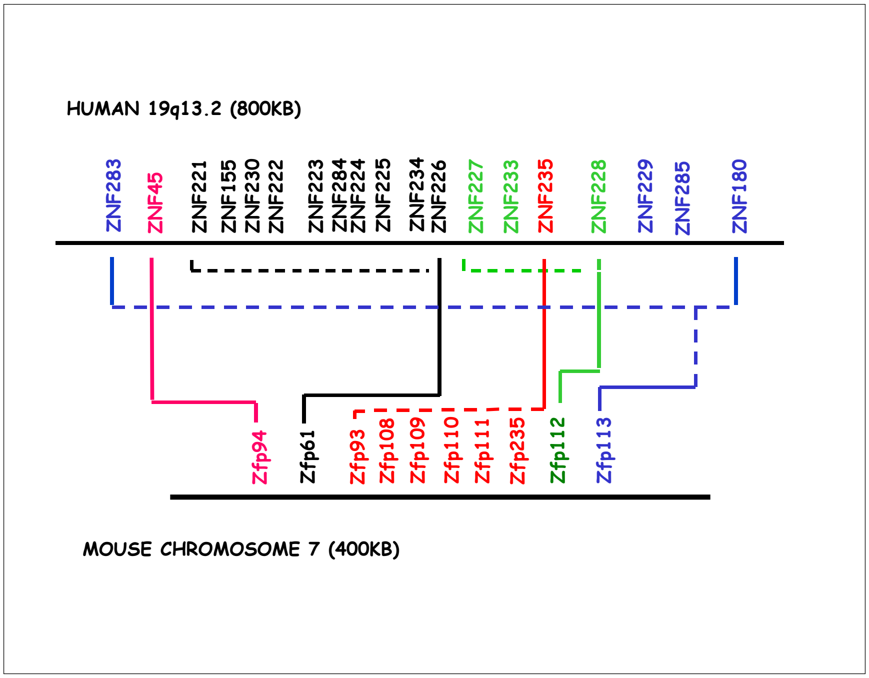
Xrcc1

znf family

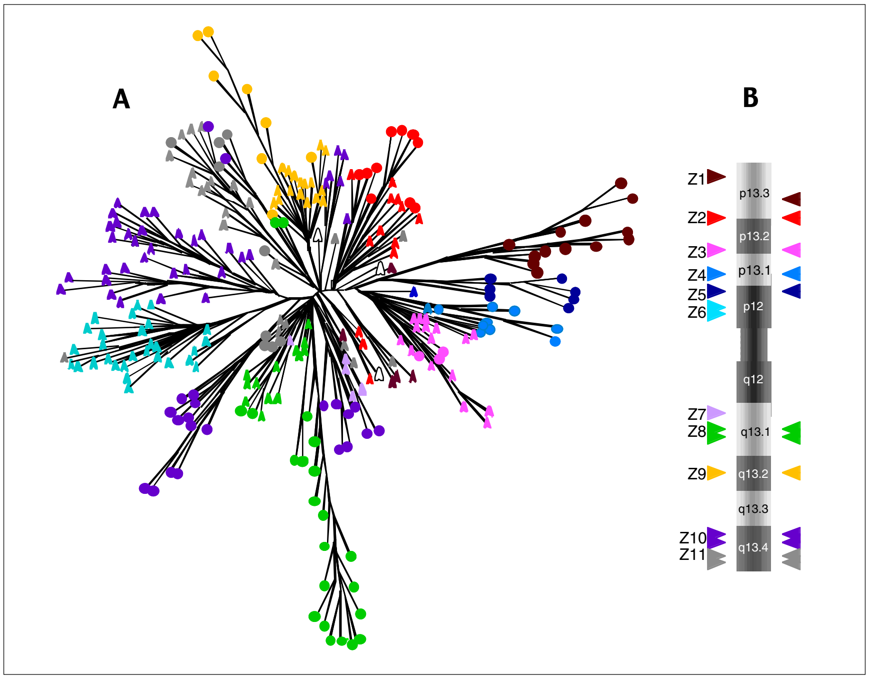
Pvr

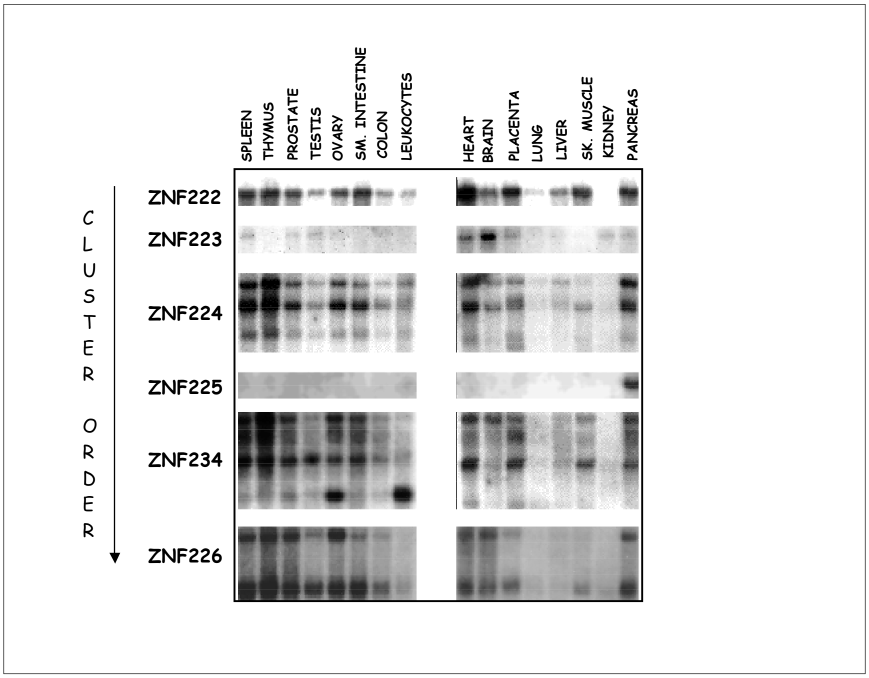
Apoe

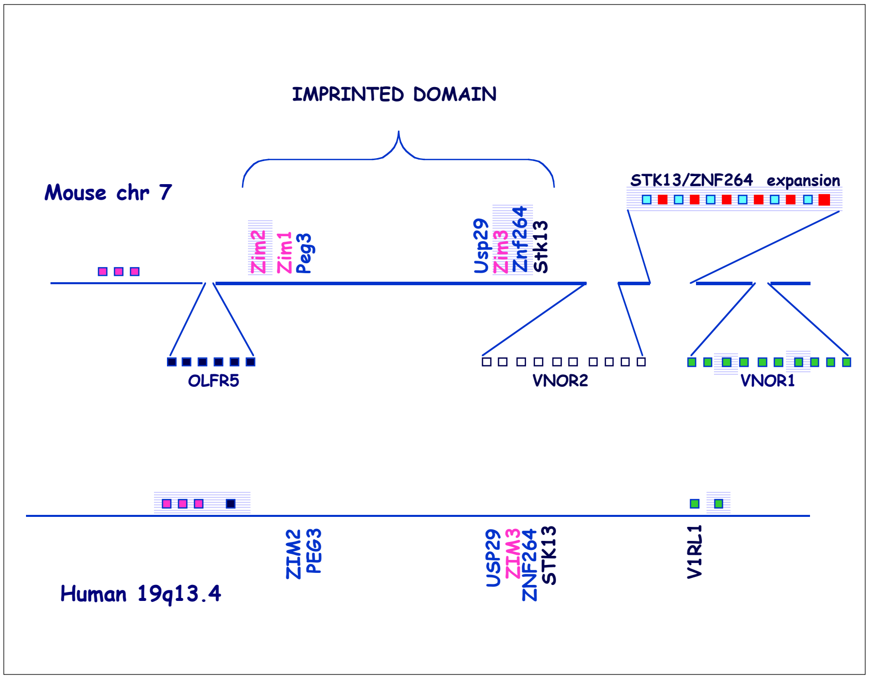
Ckmm


HUMAN 19q13.2 (800KB)

MOUSE CHROMOSOME 7 (400KB)

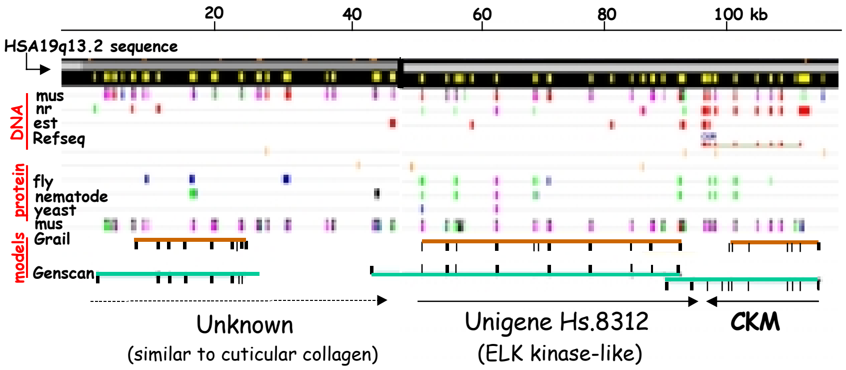



SPLEEN
THYMUS
PROSTATE
TESTIS
OVARY
SM. INTESTINE
COLON
LEUKOCYTES
HEART
BRAIN
PLACENTA
LUNG
LIVER
SK. MUSCLE
KIDNEY
PANCREAS
CLUSTER ORDER
ZNF222
ZNF223
ZNF224
ZNF225
ZNF234
ZNF226


IMPRINTED DOMAIN
Mouse chr 7
STK13/ZNF264 expansion
Zim2
Zim1
Peg3
Usp29
Zim3
Znf264
Stk13
OLFR5
VNOR2
VNOR1
ZIM2
PEG3
USP29
ZIM3
ZNF264
STK13
V1RL1
Human 19q13.4

In general...

- In contrast to unique genes regions, tandemly clustered families differ extensively in gene content, gene number and organization between the two species
  - VNO receptors, OLFR genes: multiple functional copies in mouse, and multiple pseudogenes in ch19; human singletons represented by large clusters in mouse
  - ZNF genes: conserved clusters, but different gene complements due to ongoing differential gain and loss of gene copies
  - Many actively expressed, and probably functional, lineage-specific genes exist within these and other families, at least 100 on HSA19 alone

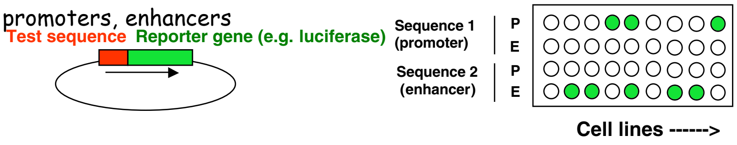


What's next?

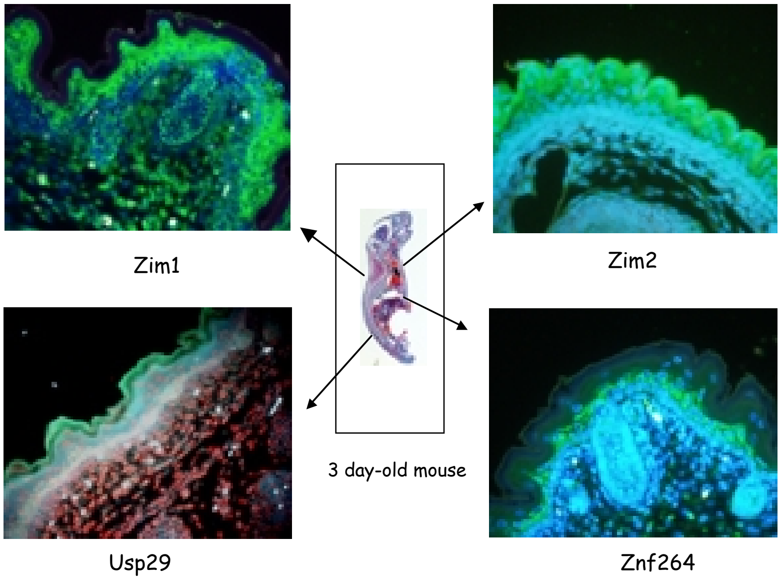
- Defining the borders of known and predicted genes
  - Which elements are linked together to create specific transcription units? Are alternative transcripts generated in different tissues?
- Identifying and testing regulatory elements predicted by comparative alignments
  - Testing function of predicted promoters, enhancers using high-throughput reporter assays
- Linking cell-type specific expression to regulatory element structure
  - can we decipher the code of gene regulation?

Triaging candidate sequences for regulatory function

- Identify candidate regions from HSA19 comparative database
  - Further computations to identify additional elements, to map locations relative to known and predicted genes, to eliminate likely exons, and supplementary evidence e.g. Maps of transcription factor binding sites
- Design oligonucleotide primers, PCR and clone putative regions into commercially available reporter-construct vectors
- Transfect candidates into arrayed cell lines and assay for activity as promoters, enhancers



Sequence 1 (promoter)	P	○ ○ ○ ● ● ○ ○ ○ ●
	E	○ ○ ○ ○ ○ ○ ○ ○ ○
Sequence 2 (enhancer)	P	○ ○ ○ ○ ○ ○ ○ ○ ○
	E	○ ● ● ○ ● ○ ● ● ○



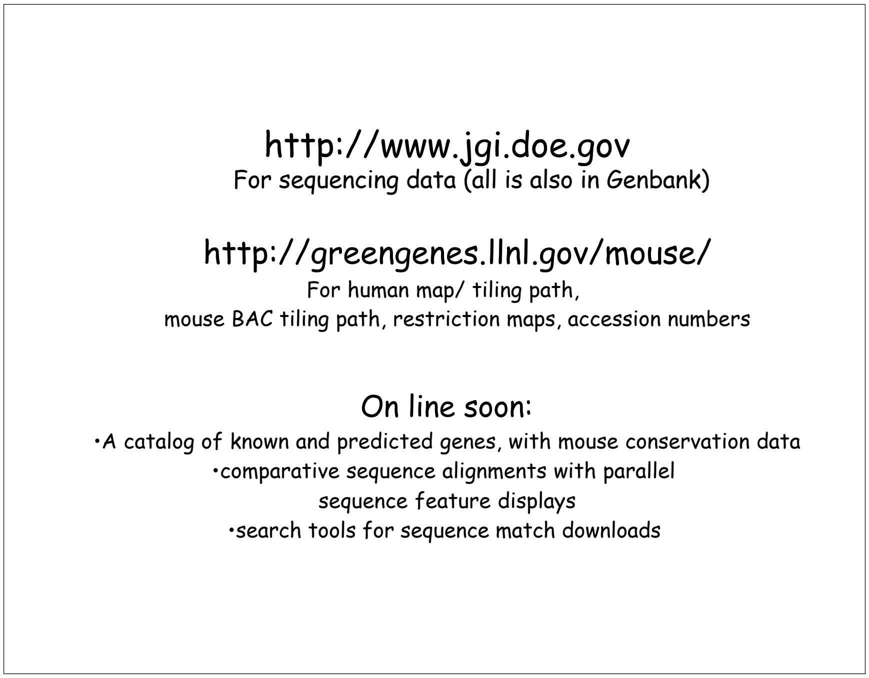
Acknowledgements

- Paramvir Dehal
- Art Kobayashi
- Anne Olsen
- Joomyeong Kim
- Laurie Gordon

- Mouse sequencing: Elbert Branscomb,Trevor Hawkins, Paul Predki, Susan Lucas, Chris Elkin, Paul Richardson, Martin Pollard & many others (JGI)

- Database design and sequence analysis: Peg Folta, Astrid Terry, Carol Zhou, Qing Zhang, Sam Rash, Dan Rokhsar (JGI); Ed Uberbacher, Miriam Land (ORNL)

- Mappers: Anne Bergmann, Hummy Badri, Mari Christensen, Chi Ha, Sha Hammond, Matt Groza, Eddie Wehri, Michelle Vargas, Mark Wagner, Mark Shannon



http://www.jgi.doe.gov

For sequencing data (all is also in Genbank)

http://greengenes.llnl.gov/mouse/

For human map/ tiling path,
mouse BAC tiling path, restriction maps, accession numbers

On line soon:

•A catalog of known and predicted genes, with mouse conservation data
•comparative sequence alignments with parallel
sequence feature displays
•search tools for sequence match downloads